View metadata, citation and similar papers at core.ac.uk
brought to you by CORE
provided by Xiamen University Institutional Repository

学校编码：10384　　分类号__密级_____
学号：23020061152441　　UDC ___

硕　士　学　位　论　文

# 动态的关联规则挖掘算法研究

# Research on Dynamic Association Rules Mining Algorithm

蓝 祺 花

指导教师姓名：张 德 富 副教授

专　业　名　称：计算机软件与理论

论文提交日期：2009　年　　月

论文答辩时间：2009　年　　月

学位授予日期：2009　年　　月

答辩委员会主席：_____

评　　阅　　人：_____

2009 年　月

厦门大学博硕士论文摘要库

厦门大学博硕士论文摘要库

# 厦门大学学位论文原创性声明

本人呈交的学位论文是本人在导师指导下,独立完成的研究成果。本人在论文写作中参考其他个人或集体已经发表的研究成果，均在文中以适当方式明确标明，并符合法律规范和《厦门大学研究生学术活动规范（试行）》。

另外，该学位论文为（　　）课题（组）的研究成果，获得（　　）课题（组）经费或实验室的资助，在（　　）实验室完成。（请在以上括号内填写课题或课题组负责人或实验室名称，未有此项声明内容的，可以不作特别声明。）

声明人（签名）：

年　　月　　日

厦门大学博硕士论文摘要库

厦门大学博硕士论文摘要库

# 厦门大学学位论文著作权使用声明

本人同意厦门大学根据《中华人民共和国学位条例暂行实施办法》等规定保留和使用此学位论文，并向主管部门或其指定机构送交学位论文（包括纸质版和电子版），允许学位论文进入厦门大学图书馆及其数据库被查阅、借阅。本人同意厦门大学将学位论文加入全国博士、硕士学位论文共建单位数据库进行检索，将学位论文的标题和摘要汇编出版，采用影印、缩印或者其它方式合理复制学位论文。

本学位论文属于：

（　　）1.经厦门大学保密委员会审查核定的保密学位论文，于　　年　　月　　日解密，解密后适用上述授权。

（√　）2.不保密，适用上述授权。

（请在以上相应括号内打“√”或填上相应内容。保密学位论文应是已经厦门大学保密委员会审定过的学位论文，未经厦门大学保密委员会审定的学位论文均为公开学位论文。此声明栏不填写的，默认为公开学位论文，均适用上述授权。）

声明人（签名）：

年　　月　　日

厦门大学博硕士论文摘要库

厦门大学博硕士论文摘要库

# 摘 要

关联规则是数据挖掘的一个重要研究方向，用于寻找数据库中不同项或属性间的相关性。而在关联规则中最常使用的方法为 Apriori 算法，但其存在一些缺点，例如，产生大量的候选项集，需多次扫描数据库等，从而造成 Apriori 算法的计算效率比较低。后续虽有许多研究针对这些缺点做改进，但大多未脱离 Apriori 算法的整体框架，以致其执行效率并无很大的改进。除此以外，在关联规则挖掘中还普遍存在两个问题：1.用户在挖掘关联规则的过程中需要预先设置挖掘参数来获取想要的规则，但往往需要通过多次调整这些参数才能达到预期的目的，如何在参数多次调整中进行高效的挖掘？2.当挖掘的数据库不断更新时，如何高效及时地挖掘出有趣的关联规则？传统的关联规则挖掘算法大都是静态的，对于上述问题，需要对数据库重新进行挖掘，执行效率不够高。

本文首先针对第一个问题，提出一个新的关联规则挖掘算法，称为 EUF 算法。该算法通过一个子集模组的对映，至多扫描两遍数据库即可计算出出现在数据库中的所有项目集的支持数，最后再由使用者输入最小支持度阈值和置信度阈值产生关联规则，算法效率不受支持度大小的影响，并且在支持度调整时无需重新扫描数据库，所以执行效率平稳快速。针对数据库更新的情况，本文在 EUF 算法的基础上提出 EUF-IU 增量更新算法，不论数据库怎么变动，只需要扫描变动的那部分数据，即可挖掘出数据库更新后的关联规则，节省了时间耗费。实验证明，本文提出的两个算法能随着参数和数据库的变化，进行动态关联规则挖掘，两个算法的执行效率在支持度较小时优于传统的关联规则挖掘算法，尤其优于 QDT 算法。

**关键词**：数据挖掘；关联规则；增量更新

厦门大学博硕士论文摘要库

厦门大学博硕士论文摘要库

# Abstract

Association rules mining is one of the significant issues in Data Mining, which describes potential relationships among data items in database.The Apriori algorithm is one of the most frequently used algorithms. But the Apriori algorithm has some defects, for example, it will produce large amounts of candidate itemsets and need scanning whole database frequently. Many researches try to improve the performance of the Apriori algorithm, but still can't escape from the frame of the Apriori algorithm mostly and lead to a little improvement of the performance. In addition, there are two prevalent problems in association rules mining: First, usually, it's necessary to set some parameters for customers before mining, and mostly they have to adjust these parameters many times to acquire the satisfactory rules, then how to implement efficiently during the repetitious process? Second, how to acquire the desired results efficiently and immediately when the mining data updates constantly? The traditional association rules mining algorithms are static, for the above problems, they must re-process the whole database again to make sure the consistence between association rules and data, result in dissatisfactory efficiency.

To solve the first problem, a new associate rules mining algorithm, named EUF, is proposed in this paper. The times that the EUF algorithm scans the transaction database needn't more than twice to calculate the support counts of all candidate itemsets by using of subset temple mapping. Then, input the support thresholds and confidence thresholds to generate the associate rules by user. The efficiency of EUF is independent of the support thresholds, and doesn't need re-process the whole database when the support threshold is adjusted. So the efficiency of EUF is steady and efficient. To solve the second problem, we propose an algorithm, named EUF-IU, based on EUF. Regardless of how the database changes, the EUF-IU only processes the modified part in the database, instead of the whole database, and then get accurate and complete association rules.Therefore, it can save the cost of runtime. Experiments show that the two algorithms presented by this paper can dynamically mine the

厦门大学博硕士论文摘要库

association rules with the changing of parameters and database, and the efficiency of the two algorithms is superior to the traditional association rules mining algorithms when the minsupports are small，especially better than the QDT algorithm.

**Key words**: Data Mining；Association Rules；Incremental Updating

厦门大学博硕士论文摘要库

# 目 录



厦门大学博硕士论文摘要库

厦门大学博硕士论文摘要库

# Contents



厦门大学博硕士论文摘要库

厦门大学博硕士论文摘要库

# 第一章 绪论

## 1.1 本课题研究背景与动机

近十几年，随着科学技术飞速的发展，经济和社会都取得了极大的进步，与此同时，在各个领域产生了大量的数据，如人类对太空的探索，银行每天的巨额交易数据。显然在这些数据中包含着丰富的信息，如何处理这些数据得到有价值的信息，人们为此进行了有益的探索。计算机技术的迅速发展使得处理海量数据成为可能，这就推动了数据库技术的极大发展，但是面对不断增加如潮水般的数据，人们不再满足于数据库的查询功能，提出了深层次问题：能不能从数据中提取信息或者知识为决策服务？就数据库技术而言已经显得无能为力了，同样，传统的统计技术也面临了极大的挑战。这就急需有新的方法来处理这些海量的数据。在这种情况下，数据挖掘这种新型的数据分析技术诞生了，它是一门交叉性学科，融合了数据库技术、人工智能、机器学习、统计学、知识工程、模式识别、信息检索、高性能计算以及数据可视化等最新技术的研究成果。

数据挖掘技术通过对海量数据的分析，发现数据之间的潜在联系，为人们提供自动决策支持，为决策者提供重要的、有价值的信息和知识，从而产生不可估量的效益。近年来，数据挖掘技术蓬勃发展，许多数据挖掘的方法被成功地提出来，这些重要技术可以归类为以下等主要研究议题：

1．关联规则（Association rule）

2．分类（Classification）

3．聚类（Clustering）

4．序列模式（Sequential pattern）

目前这些重要技术已普遍应用在银行保险、市场分析、财务分析、电子商务等各个领域[1,2,3,4]。

其中关联规则为数据挖掘中最广为讨论和应用的方法[5]，用于挖掘出隐藏在数据库中某些商品项目会引发其它商品项目出现的规则，称之为关联规则。关联规则经常被应用于市场销售中，从交易数据库中找出消费者所购买的商品项目之间的关联，从分析顾客的交易记录中，了解不同商品间的关联，挖掘出各个不

厦门大学博硕士论文摘要库

同商品间隐含的销售组合，再将这些信息运用于销售策略上，提升商品销售量，故又称为购物篮分析（Market-basket Analysis）[6]。其中最有名的例子莫过于美国的沃尔玛超市利用数据挖掘技术，发现顾客购买了尿布的同时也购买了啤酒，因此将卖场中的尿布与啤酒摆在同一架上，结果尿布与啤酒的销售量双双增长[5]，超乎卖场人员的想象。近年来随着信息化的发展，电子商务蓬勃发展，客户关系管理也应运而生，有更多的客户资料记录在数据库中，因此数据挖掘的技术显得格外重要。

自从 Agrawal[7]等人提出关联规则概念后，国内外学者对其进行了广泛深入的研究，这些研究主要涉及以下两个方面：一是算法方面，这方面包括新算法的提出和对以往算法缺陷的改进；另一方面是对关联规则概念的进一步拓宽和延伸。目前，关联规则的发现问题已经受到许多学者的极大关注，多种快速发现关联规则的挖掘算法相继被提出，其中 Agrawal 提出的 Apriori[7]算法是最普遍用来挖掘关联规则的算法，以循序渐进的方式由下而上（Bottom-Up）多次扫描数据库，计算候选项目集（Candidate Itemset）出现次数，逐步产生频繁项目集（Frequent Itemset）。然而 Apriori 算法在挖掘频繁项目集时容易产生大量的候选项目集，由于产生过多的候选项目集，Apriori 算法需花更多时间在扫描数据库上，导致 Apriori 在大型的数据库或低参数阈值的挖掘上，耗费相当多的时间，使得效率低下。因此，有不少学者从以下两方面着手改善 Apriori 算法，提高数据挖掘的效率[8,9,10,11,12,13]:

1．减少数据库扫描次数

2．降低候选项目集数量（特别是长度为 2 的候选项目集）

但是 Apriori 算法以及后来在其基础上改进和优化的算法都是静态的，基于两个共同的前提：（1）最小支持度固定不变；（2）事务数据库中的记录和项目也保持不变。然而，在日常的挖掘过程中，数据库中的数据以及挖掘的特定要求更新频繁的环境下，上述两个前提很难成立，最小支持度的调整和数据库的更新都是无法避免的[14]。因为在关联规则的实际挖掘过程中，一方面使用者为了得到适当的规则，需要反复的修改最小支持度，改善挖掘的结果，以免定的太小产生过多的规则，定的太高错失了重要的规则[15,16]。另一方面数据库的数据是在不断的被添加、修改和删除，这是一个动态的交互过程。面对这两种情况，Apriori

厦门大学博硕士论文摘要库

Degree papers are in the “Xiamen University Electronic Theses and Dissertations Database”. Full texts are available in the following ways:
1. If your library is a CALIS member libraries, please log on http://etd.calis.edu.cn/ and submit requests online, or consult the interlibrary loan department in your library.
2. For users of non-CALIS member libraries, please mail to etd@xmu.edu.cn for delivery details.

厦门大学博硕士论文摘要库